IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 12, 2020, accepted November 22, 2020, date of publication November 30, 2020,
date of current version December 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041330

Effective Video Frame Acquisition for Image Stitching

ZHE ZHANG, WANLI XUE, WEI HUANG, AND SHENGYONG CHEN, (Senior Member, IEEE)
Key Laboratory of Computer Vision and Systems (Ministry of Education), Engineering Research Center of Learning-Based Intelligent System (Ministry of Education), School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China

Corresponding authors: Wanli Xue (xuewanli@email.tjut.edu.cn) and Wei Huang (weihuang@tjut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62020106004 and Grant 61906135; in part by the Tianjin Science and Technology Plan Project under Grant 20JCQNJC01350; and in part by the National Key Research and Development Program of China under Grant 2018YFB1305200.

ABSTRACT We present an effective video frame (including reference frame and key frames) acquisition method for image stitching. The method simultaneously analyzes different types of factors, namely, the video-level stability, image-level stability, and content scale stability, to take advantage of their complementary strengths. We model the three factors with three modules that are learned from an analysis of the shooting process. The video stabilization module (VSM) selects a stable segment, while the shooting distance module (SDM) obtains a similar content scale. They collaborate during the reference video sequence so that they can benefit from each other. Then, the image quality module (IQM) obtains a reference frame from the above sequence by choosing high-quality images. Finally, to obtain the key frame set, the SDM and IQM are again used to continuously filter the overlapping video sequences formed by the reference frame or the latest key frame. In particular, a comprehensive dataset containing a variety of challenges and scenarios is introduced. We have conducted an extensive set of experiments on this dataset. The results confirm the effectiveness of each module and their collaboration; our method outperforms current state-of-the-art methods.

INDEX TERMS Image stitching, effective video frame, image stitching dataset.

I. INTRODUCTION

Image stitching is the study of combining a group of images to form a single wider field of view (FOV) image [1]. These images need to have as little parallax as possible, good image quality, similar content scale, and certain overlap rate. However, for different reasons, some scenes cannot be shot to meet these requirements; in these cases, image stitching must be performed through video. For example, we want to create a panoramic image of a campus. In this scene, we cannot take images one by one and we cannot guarantee that every image we take meets the requirements of stitching. Typically, unmanned aerial vehicle (UAV) is used to shoot video around the campus and video frames that are suitable for stitching are captured from the video. Image stitching from videos encounters some challenges due to flaws in the shooting process, and a simple selection of images can create some problems in the stitching results. Therefore, to obtain stable stitching results, it is necessary to select effective video frames (EVFs) that meet the requirements of stitching.

There are defects that occur during the shooting process. First, video-level instability: the parallax is directly caused by changes in the shooting angle and path. Second, the image-level instability: on the one hand, vibrations, instability control, and rapid movement of a shooting device can cause an image to blur; on the other hand, the motion of an object can bring about motion blur. Finally, the scales of the shooting content are different: taking an aerial image as an example, different flying heights usually cause inconsistent subject sizes. Therefore, although good progress has been made in image stitching, the above disadvantages still affect the stitching results. An effective selection of video frames can avoid the above problems.

Therefore, it is essential to obtain EVFs with good stability, excellent image quality, and uniform scales for image stitching. For video-level instability, we find that a bundled camera path [2](we define bundled camera paths as spatially varying camera paths) can describe an instability well, and a video sequence with a small change in the bundled path

The associate editor coordinating the review of this manuscript and approving it for publication was Jinjia Zhou.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

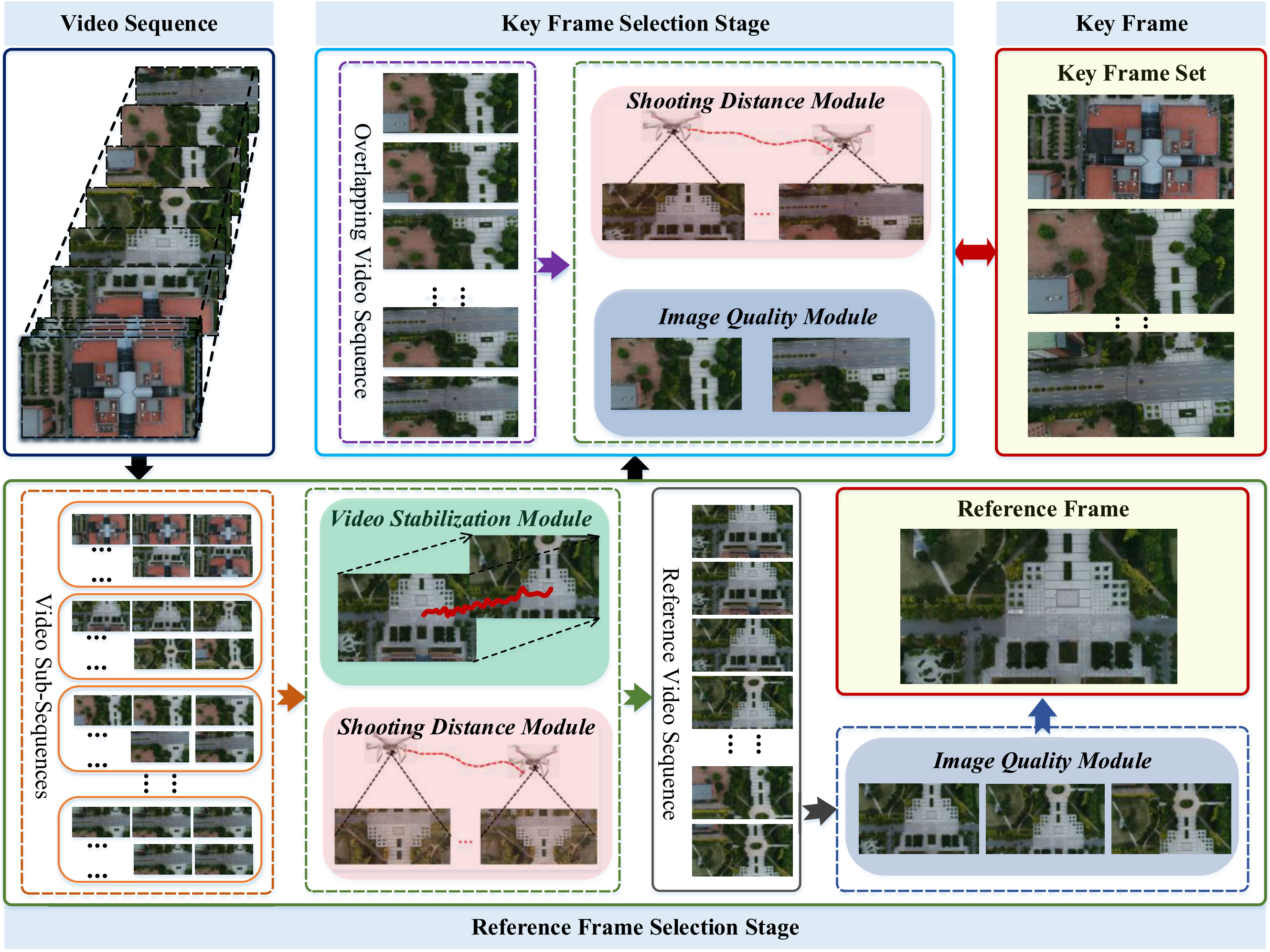

FIGURE 1. The overview of the proposed method.

has excellent stability. For image-level instability, we apply the principle that the original high-frequency content of a blurred image is lost [3]. The more high-frequency content of an original image is lost, the more blurred the image is. The main reason content is shot with inconsistent scales is that the distance, called the camera-scene (C-S) distance, between the camera lens and the scene changes; hence, it is important to obtain images with similar C-S distances.

Based on the above analysis, we have specifically proposed three modules, namely, the video stabilization module (VSM), image quality module (IQM), and shooting distance module (SDM), to address these three issues separately. Specifically, the VSM uses a "warping-based motion model" to solve the problem of selecting a stable video segment. The IQM utilizes a "no-reference perceptual blur metric" to handle the matter of choosing high-quality images. The SDM obtains a uniform C-S distance through a simple geometric calculation to cope with the issue of selecting similar content scales. The three modules work together to select EVFs for stitching.

Furthermore, an EVFs include a reference frame (the most stable video frame in the video sequence that is also the basis for selecting key frames) and key frames (frames selected from the local series of ordinary frames to represent the local frame and to record the local information). In image stitching, all images are projected onto a reference plane. Generally, taking the first image or the middle image as the reference plane, if the reference image happens to be unstable, it may visually affect the naturalness of the panorama. Because the reference frame is the most stable video frame in a video sequence, we use the reference frame as the reference plane. In the process of reference frame selection, to select a stable reference frame, a stable video segment needs to be selected first; hence, we carry out video processing based on feature point trajectory analysis to segment the complete video sequence into several video subsequences with stable backgrounds. An overview of the proposed method is illustrated in Fig. 1.

In particular, we find that the current datasets used in image stitching have some shortcomings. On the one hand, they lack pertinence, as most datasets are used for visual object tracking; on the other hand, they only consider single challenges, such as only considering the scene diversity. Therefore, we propose a new dataset for image stitching that has a total of 32 video sequences, which include challenges that arise from various changes in flying height, flying speed, and video stability. We perform a series of experiments on this dataset, and the experimental results not only show the superiority of the proposed method but also verify the validity of the dataset.

In this paper, a novel EVF acquisition-based image stitching method is proposed. Different from the typical image stitching methods, the proposed method conducts stitching with a video sequence as the carrier. First, EVFs are obtained from a video to meet the needs of image stitching. On the basis of meeting the requirements of the overlap rate, we also comprehensively evaluate the video-level stability, the image-level stability and the content scale stability and select EVFs to improve the stitching performance. Furthermore, the EVFs are divided into reference frame and key frames. On the one hand, the reference frame is used as the basis for selecting the key frames. On the other hand, the reference frame is the most stable video frame in the video. We use it as the reference plane of projection during stitching, which can improve the naturalness of stitching.

The contributions of this paper mainly include three aspects:

(1) A image stitching framework based on effective video frame acquisition is proposed, which can realize end-to-end image stitching with a video as the carrier.

(2) A novel effective video frame acquisition method is proposed. Based on a comprehensive evaluation of video-level stability, image-level stability, content scale stability and overlap rate, effective video frames are selected and divided into a reference frame and key frames. The VSM, IQM, and SDM are proposed to address different problems caused by the shooting process;

(3) A comprehensive dataset containing a variety of challenges and scenarios is proposed.

II. RELATED WORKS

A. IMAGE STITCHING

The typical image stitching method usually uses a global transform (such as affine, similarity and projection) to register the overlapped areas of an image, we call this method Homography. Brown and Lowe [4] proposed AutoStitch algorithm. Similar to Homography, a global transformation was used, and a bundle adjustment was used to calculate image coordinate transformation parameters. Then, to deal with parallax and improve the registration accuracy. Gao *et al.* [5] proposed a dual homography method that blends the homography estimated for the distant plane with the homography estimated for the ground plane adaptively according to the positions of feature points. Zaragoza *et al.* [6] proposed the as projective as possible (APAP) algorithm, which effectively improves the registration accuracy for large parallax images by multiple homographies. Lin *et al.* [7] improved the stitching performance gradually by using an iterative warp and seam estimation. Lee and Sim [8] proposed a video stitching algorithm for a large parallax based on epipolar geometry. Lee *et al.* [9] proposed an image mosaic algorithm with robustness to large disparities based on the new concept of warping residuals.

Through various registration methods, the overlapping areas of two images can be well aligned and the nonoverlapping areas usually have serious distortion. The shape preserving half projection (SPHP) algorithm [10] was proposed; it corrects the shape of the stitched image and reduces the projection distortion. Lin *et al.* proposed a homographic linearization method [11], which is also a shape correction problem, and the natural appearance of the stitching results is improved compared with the natural appearence of the results of SPHP. Chen *et al.* [12] estimated the proper scale and rotation for each image and designed an objective function for warping estimation based on a global similarity prior. Li *et al.* [13] proposed a novel quasi homography to solve the line blending problem between the homography transformation and the similarity transformation by linearly scaling the horizontal component of the homography to create a more natural panorama. In 2019, [14] presented an illumination-smoothing image stitching method based on the shape-optimizing hybrid transformation. The single perspective warps (SPW) algorithm [15] applies two single-perspective warps for natural image stitching.

Image stitching has been well developed, especially in image registration. However, in image stitching with a video as the carrier, it is not enough only to apply the existing technology for stitching because a video has redundancy and some defects in the shooting process, which lead to stitching failures. Furthermore, the proposed EVF acquisition method is used to obtain the images meeting the requirements of image stitching. Then, multiple images are spliced together. In the image registration stage, we apply the existing AutoStitch algorithm.

B. EFFECTIVE VIDEO FRAME ACQUISITION

Currently, the most common EVF acquisition method is based on the fixed interval method. For example, Yang *et al.* [16] used a fixed time interval (every two seconds) to extract video frames as key frames. This method can solve the problem of video redundancy to a certain extent but cannot guarantee a constant overlap rate between frames. The most basic requirement of images used in stitching is that a certain overlap rate should be met between images. To ensure a constant overlap rate, some new EVF acquisition methods were proposed. Bang *et al.* [17] focused on preprocessing in image stitching. By understanding the height and speed of a UAV, the triangulation principle is utilized to choose key frames with a certain overlap between the images. Dhanda *et al.* [18] proposed a method to analyze the overlap between images and filter out images through image metadata when analyzing the aerial data of UAVs to reduce video redundancy and inconsistencies. Bu *et al.* [19] employed monocular simultaneous localization and mapping (SLAM) to perform real-time stitching based on UAV images. During the selection of EVFs, they calculated the relative distance between two frames through the weighted combination of translation and rotation in large scale direct SLAM (LSD-SLAM). The key frames were selected by judging the relationship between the relative distance and threshold.

Although these methods can ensure a constant overlap rate among frames, they fail to take into account some important factors that affect the performance of the panorama, such as video stability, image quality, and image content scale, as shown in Fig. 2(a). Moreover, these methods all use the first frame in the video as the reference frame and then select the key frames. When the image quality of the first frame is poor, it will lead to catastrophic consequences for the stitching results, as shown in Fig. 2(b).

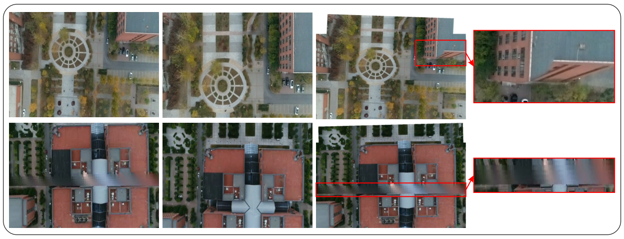

FIGURE 2. The challenges of effective video frames in image stitching. Images of the first and second columns are stitched, and we show results on the third column. Top: (a) The challenge of large scale at the level of image content in image stitching. Bottom: (b) The challenge of the reference-frame quality in image stitching.

C. THE DATASET

In addition, the datasets used in image stitching research are mainly derived from public datasets and datasets created by authors. Literature [20] introduced an efficient stitching system and experimented on the publicly available VIVID dataset [21]. In literature [16], a valid graph-based framework stitching method is presented, and VIRAT benchmark aerial video dataset [22] is used. The SkyStitch algorithm proposed by Meng *et al.* [23] in 2015 provides users with a panoramic video stream by stitching together multiple aerial video streams. The data come from a drone video taken by the author. Bang *et al.* [17] attempted to select EVF parts to create high-quality panoramas. The experimental data were derived from the author's aerial videos but not disclosed. In 2016, Bu *et al.* [19] developed the NPU DroneMap Dataset, which includes original data consisting of videos, flight logs, GCPs, and camera calibration data.

By analyzing the data sources in these articles, it is found that the VIVID dataset is a tracking dataset proposed by Robert T. Collins *et al.* The VIRAT dataset, initially provided by Defense Advanced Research Projects Agency (DARPA), is used for video surveillance. The challenges faced by the NPU DroneMap Dataset are not comprehensive, the classification is not precise, and there is no flight control information. These datasets either lack challenges or lack scene types.

III. THE PROPOSED APPROACH

Our proposed method is EVF acquisition, which improves the performance of the stitching results. An overview of the method is described in Fig. 1. The whole process is divided into two stages: the reference frame selection stage and the key frame selection stage. There are three main functional modules (VSM, IQM and SDM) that address video-level stability, image-level stability and content scale stability. The VSM estimates the stability of the video subsequence. Inspired by Liu *et al.* [2], we use bundled camera paths to evaluate the stability of a video segment. Additionally, a video sequence is first divided into several video subsequences with stable backgrounds. The IQM calculates the blur value of a video frame. We use the ‘‘no-reference perceptual blur metric’’ method to obtain the blur value. The SDM helps to select EVFs with small differences in image content scale. This chapter describes each module and provides an overview of the proposed method in detail.

A. VIDEO PREPROCESSING

In the reference frame selection stage, to find a stable reference frame, since it is impossible to calculate whether a video frame is stable in the video, we find a stable video segment and select the reference frame from the stable video segment to ensure that the reference frame is stable. Here, we divide the video into several video segments based on feature point trajectory analysis. A schematic diagram is shown in Fig. 3; V_i represents the i-th video segment.

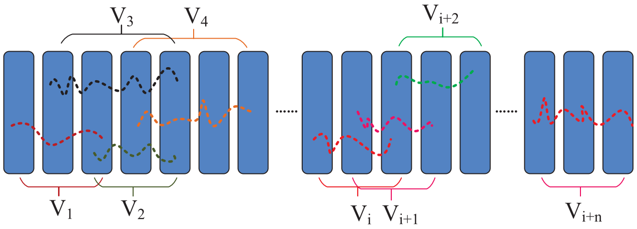

FIGURE 3. Video segmentation diagram. A dotted line segment represents a trajectory. A rectangular box represents a video frame.

In this paper, a standard Kanade-Lucas-Tomasi (KLT) tracker is used to track feature points and depict motion trajectories, and each motion trajectory is a video segment. The KLT algorithm also performs well in tracking, especially in real-time computing. Because the video is dynamic, the moving foreground may appear in the shot content, which affects the segmentation results. In addition, considering the vigorous motion of objects and cameras, the exposed part of the background is constantly changing, which makes background tracking impossible. Therefore, to make our segmentation more robust, we need to use features from the background region to remove the foreground interference that may be generated in the video. Then, we use the robust background identification method [24], which can reliably identify background features in complicated videos, allowing us to perform our work only on the background area, thereby avoiding the negative impact of the foreground features.

B. VIDEO STABILIZATION MODULE

A video may be unstable if the shooting angle or path changes during the shooting process. When we select a reference frame, we want to select it in a stable video segment. If the reference frame is in an unstable video segment, it is possible that there will be a large parallax between the reference frame and the key frame, and the stitching results may be unnatural due to the large parallax. To ensure that the selected reference frame is in a stable video segment, we use the idea of bundled camera paths to calculate the stability of each video segment. We use an image projective transform model to represent motion between successive video frames. Based on the proposed motion estimate model, we construct a bunch of camera paths. Each camera path is a cascade of projective transform models at each frame over time. By estimating the projective transform model, we can define a spatially varying camera path for each video subsequence.

Suppose that the reference image and target image are denoted as I and I', respectively. Given a correspondence (p, p'), p is a point in I and p' is a point in I', where $p = [x\, y]^T$ and $p' = \left[x'\, y'\right]^T$. A projective warp transforms p to p' following the relation.

$$\tilde{p}' \sim H\tilde{p} \tag{1}$$

The homogeneous coordinates of p and p' are $\tilde{p} = [x\, y\, 1]^T$ and $\tilde{p}' = \left[x'\, y'\, 1\right]^T$, respectively. $\sim$ indicates equality up to scale. A projective transform model $H \in R^{3\times 3}$ is given.

$$H = \begin{bmatrix} h1 & h2 & h3 \\ h4 & h5 & h6 \\ h7 & h8 & h9 \end{bmatrix} \tag{2}$$

Let $F_i(t)$ be the projective transform model estimated from the tth frame to the $t-1$th frame in the video segment V_i. Additionally,

$$F_i(1) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{3}$$

Let $P_i(t)$ be the bundled camera path of the video segment V_i. It can be written as:

$$P_i(t) = P_i(t-1)F_i(t) \tag{4}$$

$$P_i(t-1) = F_i(1)\cdots F_i(t-1) \tag{5}$$

$$E_s(i) = \sum_{i=1}^{3}\sum_{j=1}^{3} a_{ij} \tag{6}$$

where a_{ij} is the element in the ith row and jth column of $P_i(t)$, and $E_s(i)$ is the stable value of video segment V_i.

C. IMAGE QUALITY MODULE

Blurred images are an essential factor that causes poor quality stitching results [4], [25], [26]. In this section, we follow the principle that a blurred image loses its original high-frequency content, and the blurriness of an image is quantified without reference to other models. Algorithm 1 shows the process for calculating image blur values.

Algorithm 1 The Process of Calculating Image Blur

```
Input: The original image $F$
Output: The blur value of image $F$
  The original image $F$ is smoothed using a Gaussian filter to get the filtered image $B$.
  Get the intensity difference in the vertical and horizontal directions of image $F$.
      $D\_F_{ver}(row, col) = \|F(row, col) - F(row-1, col)\|$
      $D\_F_{hor}(row, col) = \|F(row, col) - F(row, col-1)\|$
  Get the sum of the intensity difference in the vertical and horizontal directions of image $F$.
      $S\_F_{ver} = \sum_{i=1}^{rov}\sum_{j=1}^{col} D\_F_{ver}(i,j)$
      $S\_F_{hor} = \sum_{i=1}^{row}\sum_{j=1}^{col} D\_F_{hor}(i,j)$
  Get the intensity difference in the vertical and horizontal directions of image $B$.
      $B\_F_{ver}(row, col) = \|B(row, col) - B(row-1, col)\|$
      $B\_F_{hor}(row, col) = \|B(row, col) - B(row, col-1)\|$
  Calculating the sum of the intensity of the high-frequency information lost in the vertical and horizontal directions.
      $S\_V_{ver} = \sum_{i=1}^{row}\sum_{j=1}^{col} \max\left(0, D\_F_{ver}(i,j) - D\_B_{ver}(i,j)\right)$
      $S\_V_{hor} = \sum_{i=1}^{row}\sum_{j=1}^{col} \max\left(0, D\_F_{hor}(i,j) - D\_B_{hor}(i,j)\right)$
  Calculating the proportion of high-frequency information left.
      $b\_F_{ver} = \left(S\_F_{ver} - S\_V_{ver}\right)/S\_F_{ver}$
      $b\_F_{hor} = \left(S\_F_{hor} - S\_V_{hor}\right)/S\_F_{hor}$
  The larger of $b\_F_{ver}$ and $b\_F_{hor}$ is used as the blur value.
      $blur\_F = \max\left(b\_F_{ver}, b\_F_{hor}\right)$
```

The smaller $blur\_F$ is, the more blurred the image. Here, we do not need to define a threshold value to judge whether the image is blurred or unblurred, and we only need to calculate the blurred value of an image. In the process of EVF acquisition, we comprehensively evaluate whether the video frame is suitable to be an effective frame from many aspects, rather than judge whether the video frame can be an effective frame based on the image quality alone. For ease of expression,we use $E_b(f)$ to represent the blur value of video frame f.

$$E_b(f) = blur\_F \tag{7}$$

D. SHOOTING DISTANCE MODULE

The SDM can address the challenge of unstable content scale, that is, inconsistent content scale. The main reason for the inconsistent content scale is that the C-S distance between the

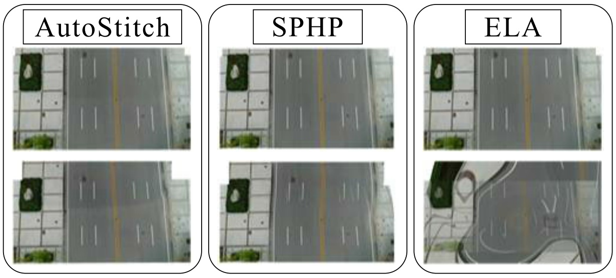

FIGURE 4. The challenges of the shooting distance in image stitching.Up: (a)The shooting distance is 34.6m and 34.9m.Bottom: (b)The shooting distance is 34.6m and 28.7m.

camera lens and the scene changes. The larger the differences in the C-S distances are, the larger the difference in the scale of the image content is. If images with large differences in content scale are stitched, it may result in unsuccessful stitching or misalignment. As shown in Fig. 4, Fig. 4(a) uses two images with C-S distances of 34.6 m and 34.9 m and applies the AutoStitch [4], SPHP [10], and ELA [27] algorithms to obtain good results. Fig. 4(b) uses two images with C-S distances of 34.6 m and 28.7 m. Both AutoStitch and SPHP have different degrees of misalignment, and the stitching result of ELA is seriously distorted.

In our proposed method, one of the criteria for selecting a reference video segment is close C-S distances among the frames in the video segment. Therefore, we calculate the average distance among frames in a video segment to evaluate the stability of the video segment in terms of the shooting distance, as shown in Eq. (8).

$$E_h(i) = \frac{\sum_{t \in V_i} \|h(t) - h(t-1)\|}{len(V_i) - 1} \tag{8}$$

where $h(t)$ represents the C-S distance of the t-th frame. Our work is based on the aerial dataset, and the C-S distance, which is the flying altitude of the UAV, can be obtained from the flight control information. $len(V_i)$ is the number of frames in video segment V_i.

When selecting the key frames, we calculate the C-S distance difference between the video frame and the reference frame. The smaller the difference is, the closer the content scale between the video frame and the reference frame. We use Eq. (9) to constrain the C-S distance difference between a key frame and the reference frame to make the image content scales as similar as possible.

$$E_{ref-h}(t) = \|h(t) - h(ref)\| \quad t \in \Omega_j \tag{9}$$

where $h(ref)$ represents the C-S distance of the reference frame, Ω_j represents the video sequence that satisfies the overlap rate, and $E_{ref-h}(t)$ represents the C-S distance difference between video frame t and the reference frame.

E. EFFECTIVE VIDEO FRAME ACQUISITION

Based on the description of each module above, in this section, we introduce the selection process of EVFs. An overview of the EVF acquisition method is shown in Fig. 1. The three modules (the VSM, IQM, and SDM) cooperate to complete an effective frame selection from coarse to finely divided in two stages.

In the reference frame selection stage, first, the video is divided into several video segments, and then the video stability and content scale changes in each video segment are measured with the VSM and the SDM so that a reference video segment V_{ref} can be obtained, as shown in Eq. (10).

$$V_{ref} = \arg\min_i (E_s(i) + E_h(i)) \tag{10}$$

where $E_s(i)$ represents the stability of the video segment V_i, as shown in Eq. (6), and $E_h(i)$ represents the average C-S distance difference of the video segment V_i, as shown in Eq. (8).

Then, in the reference video segment V_{ref}, the video frame with the best image quality is selected as the reference frame F_{ref} through the IQM, as shown in Eq. (11).

$$F_{ref} = \arg\max_t (E_b(t)) \quad t \in V_{ref} \tag{11}$$

where $E_b(t)$ represents the blur value of video frame t.

In the key frame selection stage, the key frames must meet the overlap rate requirement, and we first calculate the video sequence Ω_j, which meets the overlap rate requirement. In the overlapping video sequence Ω_j, we select the key frames by evaluating the C-S distance difference between the video frame and the reference frame as well as the image quality with the help of the SDM and the IQM, as shown in Eq. (12).

$$F_{key} = \arg\min_t \left(E_{ref-h}(t) + (1 - E_b(t))\right) \quad t \in \Omega_j \tag{12}$$

where $E_{ref-h}(t)$ represents the C-S distance difference between the video frame t and the reference frame, as shown in Eq. (9).

The EVF acquisition procedure is described in Algorithm 2.

F. MULTIPLE IMAGE STITCHING

Our main work is to acquire EVFs, and multi-image stitching is an improvement to AutoStitch, called R-AutoStitch. The first step of multi-image stitching is to find the reference plane [28], [29]to which all images are projected through a basic homographic warp [30]. Generally, taking the first image or the middle image as the reference plane, if the image selected as the reference plane happens to be unstable, it may visually affect the naturalness of the panorama or the registration accuracy. An example is shown in Fig. 5(a), the green line is in the horizontal direction, and the red line is in the direction of the stand, which is tilted. To avoid this phenomenon as much as possible, we use the reference frame as the reference plane for image stitching because it is the most stable frame in the video sequence. An example is shown in Fig. 5(b), the whole scene is on a horizontal line.

TABLE 1. Compared methods and characteristics. Since ICE is a software by Microsoft, we don't know whether we have discussed three problems when selecting effective video frames, so we use "?" to represent.

Characteristics	PRPSD	ICE	AutoStitch	PS	SPHP	ELA	SPW
Input:Video	✓	✓	×	×	×	×	×
Input:EVF	×	×	✓	✓	✓	✓	✓
Select the EVF	✓	✓	×	×	×	×	×
VSM	✓	?	×	×	×	×	×
IQM	✓	?	×	×	×	×	×
SDM	✓	?	×	×	×	×	×

Algorithm 2 The Effective Video Frames Selecting

```
Input: Video sequence
Output: The effective video frames
The video is divided into $n$ video segments with the stable backgrounds $V_n$ by section 3.1
for each video segment $V_i$ do
    Computer the stable value of video and the average C-S distance difference by (8)
end for
Get a reference video segment $V_{ref}$
for each frame $t$ in the reference video segment $V_{ref}$ do
    Compute the blur value of image $t$ by section 3.3
end for
Get a reference video frame $F_{ref}$ by (11)
$F_{base} = F_{ref}$
for each frame $t$ in the video segment $\Omega_j$ that satisfies the range of overlap with $F_{base}$ do
    Get the key-frames $F_{key}$by (12)
    $F_{base} = F_{key}$
end for
```

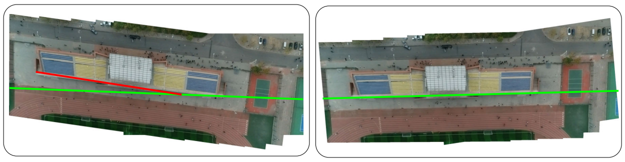

FIGURE 5. The challenges of the reference plane in multiple image stitching. Left: (a)The result of the first frame as a reference plane. Right: (b)The result of the reference frame as a reference plane.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We demonstrate the effectiveness of our proposed method in two aspects. First, we show a qualitative comparison of the stitching performance results. Second, we show a quantitative evaluation of the alignment accuracy. We also conduct ablation experiments to verify the necessity of each module. In our experiments, an aerial video dataset is used and has been made public on the website.

A. DATASET

We use DJI Phantom 4 Pro to capture videos on different terrains. The dataset has scene diversity, and it is also comprehensively challenging with variations in flying height, flying speed, video stability, and image quality. The dataset contains a total of 32 pieces of data and is publicly available on the website.

Each piece of data includes the following:

(1) The original aerial video and the converted image.

(2) The flight control file. We parse the flight control file into a CSV file.

B. QUALITATIVE COMPARISON

We compare the performance of the proposed algorithm with the following state-of-the-art algorithms: image composite editor (ICE) [31], AutoStitch [4], PhotoShop [32], SPHP [10], ELA [27], and SPW [15]. ICE is an advanced panoramic image stitcher created by the Microsoft Research Computational Photography Group. PhotoShop is a commercial tool for image processing that can complete image stitching. AutoStitch, SPHP, ELA and SPW are state-of-the-art and classical methods in the field of image stitching. The characteristics of the state-of-the-art methods are shown in Table 1, including the input and output of the algorithm, whether there is a function for selecting EVFs, and whether the three challenges are considered.

1) COMPARISON WITH ICE

In this section, we compare the proposed image stitching method based on EVF acquisition with ICE. Some stitching results are shown in Fig. 6. We choose two challenging videos for comparison, and the results show that the performance of our method is better than that of ICE, which shows serious registration errors and image quality problems (red boxes). Even though the stitching result cannot express the complete content (the ICE result of the "stand" data), the blue box indicates that the panorama is not good due to poor image quality.

2) COMPARISON OF THE EFFECTIVE VIDEO FRAME ACQUISITION METHOD

In this section, we verify the effectiveness of our selected EVF method with the stitching results. The proposed method is compared with the fixed interval method, and the results of EVF selection are verified by state-of-the-art image stitching methods (AutoStitch [4], PhotoShop [32], SPHP [10], ELA [27], and SPW [15]). The fixed interval method sets

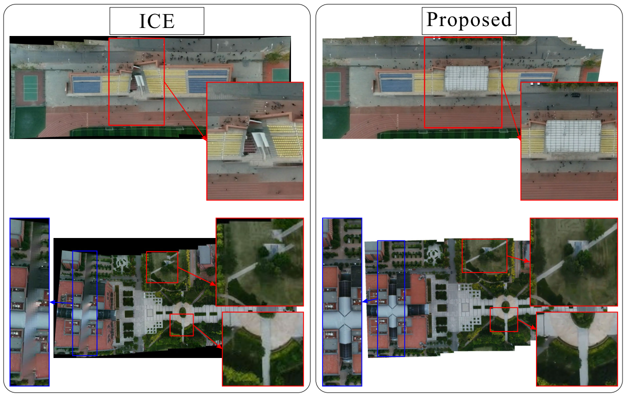

FIGURE 6. The Performance of image stitching method. Left: the results of ICE. Right: the results of our image stitching method based on effective video frame acquisition.

a fixed frame interval and fixed overlap ratio range in advance, it sets the first frame as the reference frame, and then it selects a video frame that meets the overlap rate as a keyframe.

The comparison results are shown in Fig. 7. The first row shows the stitching results of the proposed method, and the second row shows the stitching results of the fixed interval method. Each problematic region is marked with a different color box, and the same region is marked in the other result. The red box indicates the phenomenon of misalignment, the blue box shows that the poor image quality leads to an inferior panorama, and the green box shows the local distortion. The fixed interval method does not address the challenges (video-level stability, image-level stability, and content scale stability); however, the proposed method fully addresses these challenges and difficulties. It is proven that our proposed method is effective and better than the fixed interval method. Specifically, the ELA results with the fixed interval method show local distortion; however, the ELA results with the proposed method somewhat mitigate the distortion. The shape of the building in the PhotoShop result is destroyed when the fixed interval method is used; however, the shape of the building in the PhotoShop result obtained with the proposed method is presented perfectly. The results of the fixed interval method also suffer from the influence of image quality.

C. QUANTITATIVE COMPARISON

We quantitatively evaluate all the data in the dataset and compare the fixed interval method with the proposed method. The results are measured with the root mean squared error (RMSE). The RMSE is an effective parameter for evaluating registration accuracy.

$$RMSE = \frac{1}{M-1}\sum_{j=1}^{M-1}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|f\left(p_i\right)-p_i'\right\|^2} \qquad (13)$$

where $f:\mathbb{R}^2 \mapsto \mathbb{R}^2$ is a planar warp. M is the number of EVFs. N is the number of a set of point correspondences $\left\{p_i, p_i'\right\}_{i=1}^{N}$.

TABLE 2. RMSE Value Comparisons.

Methods	S/N	RMSE			
		road	parterre	stand	library
Fixed Interval	1	2.11	-	-	1.58
	2	2.24	1.84	2.71	1.95
	3	2.77	1.81	3.06	1.81
	4	2.74	2.07	2.62	1.85
	5	3.01	2.42	3.05	1.90
	6	2.92	-	-	1.95
	7	2.83	2.00	2.72	2.28
	8	-	-	2.71	1.94
PRPSD	1	1.93	1.81	2.49	1.48
	2	1.97	1.88	2.62	1.88
	3	2.73	1.80	2.78	2.00
	4	2.78	1.87	2.75	1.84
	5	2.67	2.40	2.15	1.87
	6	2.64	1.86	2.75	1.90
	7	3.02	2.05	2.64	2.27
	8	2.82	2.01	2.57	1.99

The RMSE comparison between the proposed method and the fixed interval method is shown in Tables 2. The smaller the value of the RMSE is, the better the stitching result. The red font indicates that the RMSE value is less than the corresponding value of the fixed interval method, which means that the EVFs selected by the proposed method are more suitable for mosaics than those of the fixed interval method, and the registration accuracy of the stitching result is higher. "-" indicates that the EVFs selected by the fixed interval method could not be stitched; it can be said that our proposed dataset is somewhat challenging. The values in blue font are the RMSE values of the proposed method that correspond to the "-" of the fixed interval method. Only a few RMSE values of the proposed method are higher than those of the fixed interval method, but the average difference is less than 0.1 pixels. It can be seen in the table that 18.75% of the data mosaics fail when the fixed interval method is used to select the EVFs, and 59.38% of the data show that the results using our proposed method are superior to those using the fixed interval method. This proves that the EVFs selected with our proposed method are more helpful for stitching and obtaining better registration accuracy, and the dataset we have established is comprehensive and challenging.

D. ABLATION EXPERIMENTS

The proposed EVF acquisition method takes into full account the factors of video-level stability, image-level stability and content scale stability, including the VSM, IQM and SDM. In this section, we perform ablation experiments to verify the necessity of each module. In the ablation experiment, the proposed EVF acquisition method is named New, the method with the VSM removed is named New-without-s, the method with the IQM removed is named New-without-b, and the

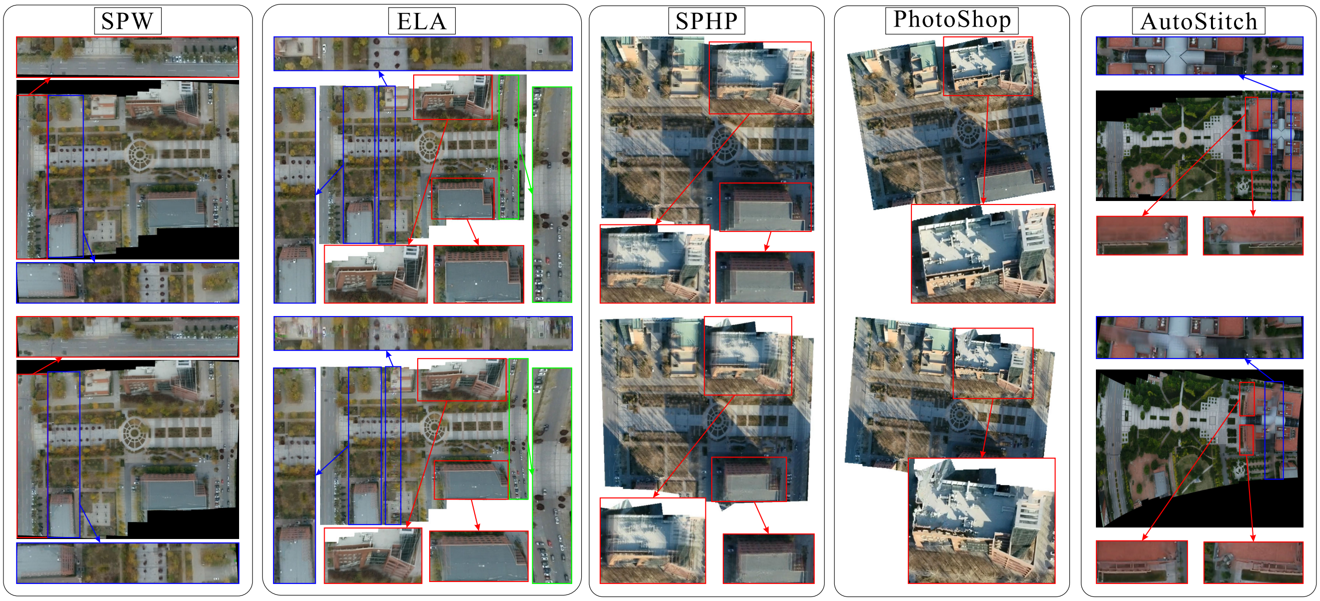

FIGURE 7. Comparisons with the fixed interval method. Up:The stitching result with our proposed method of selecting effective video frames. Bottom:The stitching result with the fixed interval method.

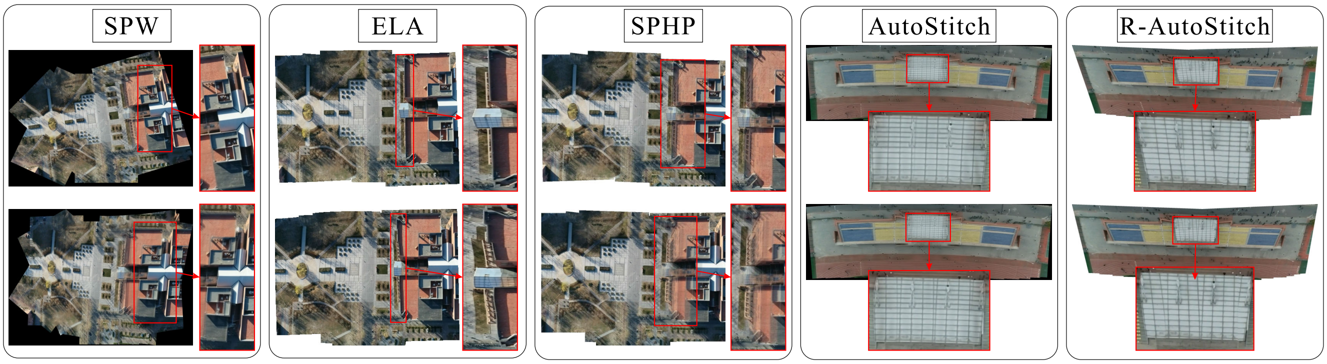

FIGURE 8. Compare the NEW method with the New-without-s method. Up:The stitching results with NEW method. Bottom:The stitching results with New-without-s method.

TABLE 3. Ablation experiments method.

Module	New	New_without_s	New_without_b	New_without_h
VSM	✓	×	✓	✓
IQM	✓	✓	×	✓
SDM	✓	✓	✓	×

method with the SDM removed is named New-without-h, as shown in Table 3.

1) NEW-WITHOUT-S

In the process of obtaining the reference frame, the New method first locates the reference video segment through the VSM and the SDM and then selects the reference frame in the reference video segment. Whereas the New-without-s method has no VSM, only the SDM is considered when selecting the reference video segment. We use Eq. (14) to locate the reference video segment; then, Eq. (11) is used to obtain the reference frame.

$$V_{ref} = \arg\min_{i} (E_h(i)) \tag{14}$$

where $E_h(i)$ represents the average C-S distance difference of video segment V_i, as shown in Eq. (8).

We compare the EVFs selected by the New method and New-without-s method on six existing image stitching methods. The comparison results are shown in Fig. 8 and Fig. 9.

In Fig. 8, the comparison results all have registration errors, which are marked with red boxes. In particular, for the ELA stitching result of the EVFs selected with the New-without-s method, the distortion is more serious. However, the ELA stitching result of the EVFs selected with the New method obtains good results. For the SPHP result of the EVFs selected

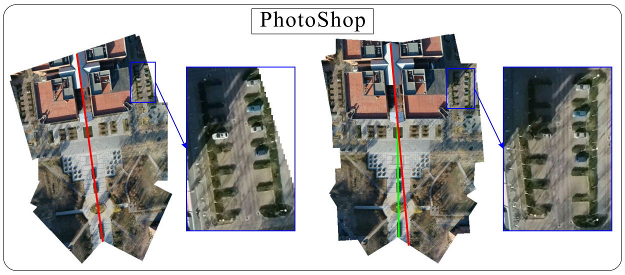

FIGURE 9. Comparison with PhotoShop. Left:The stitching results with NEW method. Right:The stitching results with New-without-s method.

with the New-without-s method, registration artifacts are generated due to inaccurate registration. The SPHP result of the EVFs selected with the New-without-s method can alleviate the problem to some extent.

Similarly, in Fig. 9, the PhotoShop result with the New-without-s method also has registration errors, which are marked with blue boxes; correspondingly, PhotoShop with the New method can obtain good stitching results. In particular, the red line represents the centerline of the building, the center of the building is in line with the center of the front square, and the green line in the right picture represents the centerline of the front square. It can be seen that the centerline of the whole scene in the right picture is inconsistent; that is to say, the whole scene is distorted. The reason for this phenomenon is the instability of the video. Therefore, the VSM plays a vital role in selecting EVFs.

2) NEW-WITHOUT-B

The New-without-b method still uses the VSM and the SDM in Eq. (10) when selecting the reference video segment. Since this method does not have the IQM, we take the first frame in the reference video segment as the reference frame, as shown in Eq. (15). When selecting the key frames, the New method comprehensively measures the image quality, the shooting distance and the overlap rate. In the New-without-b method, the frame with the smallest C-S distance from the reference frame, which satisfies a certain overlap rate, is selected as the key frame in the video segment, as shown in Eq. (16).

$$F_{ref} = V_{ref}(1) \tag{15}$$

where $V_{ref}(1)$ represents the first frame in the reference video segment.

$$F_{key} = \arg\min_{t} \left(E_{ref-h}(t)\right) \| \quad t \in \Omega_j \tag{16}$$

where $E_{ref-h}(t)$ represents the C-S distance difference between the video frame t and the reference frame, as shown in Eq. (9). Ω_j represents the video segment that satisfies the overlap rate.

We compare the effect of the New-without-b method and the New method with the two stitching methods, as shown in Fig. 10. The red box shows the inferior part in the results of ELA and PhotoShop with New-without-b. The reason for this phenomenon is due to image-level instability; there are poor quality frames in the video, and this challenge is not addressed when selecting reference frame and key frames. The New method is fully focused on this challenge, so the stitching performance obtained with the New method can yield very good results. The ELA result with New-without-b has local distortions, which are indicated by the green box. It can be seen from the comparison results in Fig. 10 that the IQM module is necessary and critical.

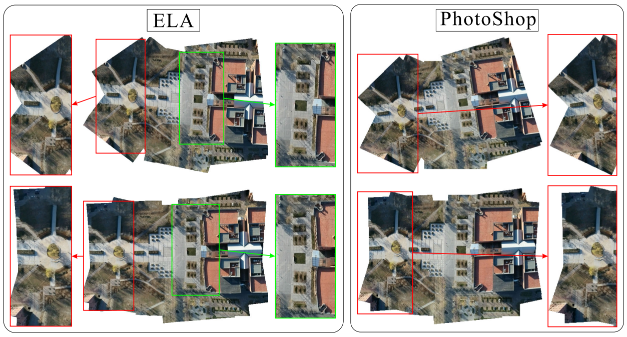

FIGURE 10. Compare the NEW method with the New-without-b method. Up: The stitching results with NEW method. Bottom: The stitching results with New-without-b method.

3) NEW-WITHOUT-H

The New-without-h method ablates the SDM, so when selecting the reference video segment, only the video stability module is considered; Eq. (17) is used to select a stable video segment.

$$V_{ref} = \arg\min_{i} \left(E_s(i)\right) \tag{17}$$

where $E_s(i)$ is a measure of the stability of the video segment V_i, as shown in Eq. (6).

When selecting key frames, it is necessary to maintain a constant overlap rate. In a video sequence Ω_j that meets the overlap rate, Eq. (18) is used to select the key frames.

$$F_{key} = \arg\max_{t} \left(E_b(t)\right) \quad t \in \Omega_j \tag{18}$$

where $E_b(t)$ represents the blur value of video frame t.

If the shooting distance is varied, the image content has different content scales, which can lead to registration errors or distortion. We test the New method against the New-without-h method with six existing stitching methods, and the comparison results are shown in Fig. 11. The ELA results with the New method and the AutoStitch results with the New method all have distortion problems, and the stitching results with the New-without-s method on the SPW, SPHP, PhotoShop and R-AutoStitch algorithms all have registration errors and artifacts. The New method performs better than the New-without-s method with all six existing

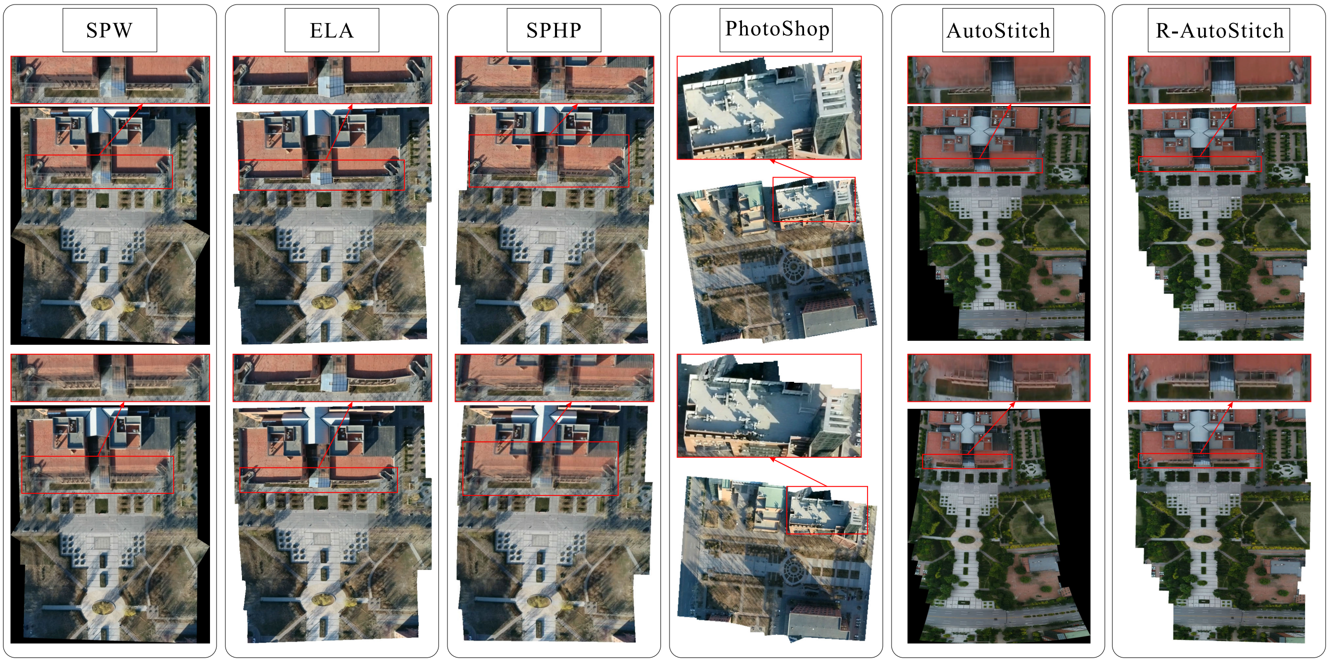

FIGURE 11. Compare the NEW method with the New-without-h method. Up: The stitching results with NEW method. Bottom: The stitching results with New-without-h method.

stitching methods. Therefore, the shooting distance module plays an important role in the process of EVF acquisition.

V. CONCLUSION

We have proposed an image stitching framework based on EVF acquisition, which is end-to-end image stitching algorithm with a video as the carrier. Specifically, we focus on the effective video frame acquisition method based on the collaboration of three modules with multifaceted stability. The modules take advantage of different levels (e.g., video, image, and content) of stability during the reference frame and key frame selection process and thus can account for most challenges in stitching. In particular, the VSM, the SDM and the IQM are used collaboratively in a reference frame selection stage, forming a collaborative reference-frame stage that is not vulnerable to image redundancy and can make the reference plane of stitching more stable. Furthermore, the SDM and the IQM are again used collaboratively to find high-quality and similar-scale images, forming the key-frame selection stage, which increases the stitching reliability. The reference frame selection stage and the key frame selection stage determine the EVFs, and an optimal frame is estimated via a novel coarse-to-fine search strategy. The experiments on the challenging dataset, which was made public, confirm that the collaboration of the three modules actually improves performance, and our method generally outperforms most existing methods.

REFERENCES

[1] T.-Z. Xiang, G.-S. Xia, X. Bai, and L. Zhang, "Image stitching by line-guided local warping with global similarity constraint," *Pattern Recognit.*, vol. 83, pp. 481–497, Nov. 2018.

[2] S. Liu, L. Yuan, P. Tan, and J. Sun, "Bundled camera paths for video stabilization," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 78:1–78:10, 2013.

[3] F. Crété-Roffet, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric," *Proc. SPIE*, vol. 12, pp. 64920I-1–64920I–11, Feb. 2007.

[4] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, Apr. 2007.

[5] J. Gao, S. J. Kim, and M. S. Brown, "Constructing image panoramas using dual-homography warping," in *Proc. CVPR*, Jun. 2011, pp. 49–56.

[6] J. Zaragoza, T.-J. Chin, Q.-H. Tran, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving DLT," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1285–1298, Jul. 2014.

[7] K. Lin, N. Jiang, L. F. Cheong, M. Do, and J. Lu, "Seagull: Seam-guided local alignment for parallax-tolerant image stitching," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 370–385.

[8] K.-Y. Lee and J.-Y. Sim, "Stitching for multi-view videos with large parallax based on adaptive pixel warping," *IEEE Access*, vol. 6, pp. 26904–26917, 2018.

[9] K.-Y. Lee and J.-Y. Sim, "Warping residual based image stitching for large parallax," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8198–8206.

[10] C.-H. Chang, Y. Sato, and Y.-Y. Chuang, "Shape-preserving half-projective warps for image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3254–3261.

[11] C.-C. Lin, S. U. Pankanti, K. N. Ramamurthy, and A. Y. Aravkin, "Adaptive as-natural-as-possible image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1155–1163.

[12] Y.-S. Chen and Y.-Y. Chuang, "Natural image stitching with the global similarity prior," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 186–201.

[13] N. Li, Y. Xu, and C. Wang, "Quasi-homography warps in image stitching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1365–1375, Jun. 2018.

[14] S. Liu and Q. Chai, "Shape-optimizing and illumination-smoothing image stitching," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 690–703, Mar. 2019.

[15] T. Liao and N. Li, "Single-perspective warps in natural image stitching," *IEEE Trans. Image Process.*, vol. 29, pp. 724–735, 2020.

[16] T. Yang, J. Li, J. Yu, S. Wang, and Y. Zhang, "Diverse scene stitching from a large-scale aerial video dataset," *Remote Sens.*, vol. 7, no. 6, pp. 6932–6949, May 2015.

[17] S. Bang, H. Kim, and H. Kim, "UAV-based automatic generation of high-resolution panorama at a construction site with a focus on preprocessing for image stitching," *Autom. Construct.*, vol. 84, pp. 70–80, Dec. 2017.

[18] A. Dhanda, F. Remondino, and M. S. Quintero, "A metadata based approach for analyzing UAV datasets for photogrammetric applications," *ISPRS-Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 297–302, May 2018.

[19] S. Bu, Y. Zhao, G. Wan, and Z. Liu, "Map2DFusion: Real-time incremental UAV image mosaicing based on monocular SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4564–4571.

[20] J. Li, T. Yang, J. Yu, Z. Lu, P. Lu, X. Jia, and W. Chen, "Fast aerial video stitching," *Int. J. Adv. Robot. Syst.*, vol. 11, no. 10, p. 167, Oct. 2014.

[21] R. T. Collins, X. Zhou, and S. K. Teh, "An open source tracking testbed and evaluation Web site," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill.*, 2005.

[22] X. Wang, E. Swears, A. Hoogs, S. Oh, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, and D. Ramanan, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2011.

[23] X. Meng, W. Wei, and B. Leong, "Skystitch: A cooperative multi-uav-based real-time video surveillance system with stitching," in *Proc. Aust. Conf. Comput. Vis. (MM)*, Oct. 2015, pp. 261–270.

[24] F.-L. Zhang, X. Wu, H.-T. Zhang, J. Wang, and S.-M. Hu, "Robust background identification for dynamic video editing," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Nov. 2016.

[25] M. R. Jahanshahi, S. F. Masri, and G. S. Sukhatme, "Multi-image stitching and scene reconstruction for evaluating defect evolution in structures:," *Struct Health Monit*, vol. 10, no. 6, pp. 643–657, 2011.

[26] J. H. Chen and C. M. Huang, "Image stitching on the unmanned air vehicle in the indoor environment," in *Proc. SICE Conf.*, 2012, pp. 402–406.

[27] J. Li, Z. Wang, S. Lai, Y. Zhai, and M. Zhang, "Parallax-tolerant image stitching based on robust elastic warping," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1672–1687, Jul. 2018.

[28] R. Marzotto, A. Fusiello, and V. Murino, "High resolution video mosaicing with global alignment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2004, p. 1.

[29] E. Kang, I. Cohen, and G. Medioni, "A graph-based global registration for 2D mosaics," in *Proc. Spain. Conf. Comput. Vis. (ICPR)*, vol. 1, Sep. 2000, pp. 257–260.

[30] J. Chen, Q. Xu, L. Luo, Y. Wang, and S. Wang, "A robust method for automatic panoramic UAV image mosaic," *Sensors*, vol. 19, no. 8, p. 1898, Apr. 2019.

[31] (2015). *Image Composite Editor*. [Online]. Available: https://www.microsoft.com/en-us/research/product/computational-photography-applications/image-composite-editor/

[32] (2019). *PhotoShop*. [Online]. Available: https://adobe-photoshop.en.softonic.com

ZHE ZHANG received the B.S. degree in computer science and technology from Hebei Normal University, in 2013, and the M.S. degree in computer technology from the Beifang University of Nationalities, in 2015. She is currently pursuing the Ph.D. degree in computer science and technology with the Tianjin University of Technology. Her research interest includes image stitching.

WANLI XUE received the B.S. degree in pure and applied mathematics from Tianjin Polytechnic University, in 2009, and the Ph.D. degree in technology of computer application from Tianjin University, in 2019. He is currently a Lecturer with the School of Computer Science and Engineering, Tianjin University of Technology. His research interests include visual tracking and images stitching.

WEI HUANG received the Ph.D. degree in optical engineering from the Institute of Modern Optics, Nankai University, Tianjin, China, in 2016. In 2016, she joined the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, as a Lecturer. Her current research interest includes inverse design and prediction of optical structure based on machine learning pattern recognition.

SHENGYONG CHEN (Senior Member, IEEE) received the Ph.D. degree in robot vision from the City University of Hong Kong, Honk Kong, in 2003. From 2006 to 2007, he was with the University of Hamburg, Hamburg, Germany. He is currently a Professor with the Tianjin University of Technology, Tianjin, China. His current research interests include computer vision, robotics, and image analysis.